И.В. Соченков

Метод сравнения текстов для решения поисково-аналитических задач[1]

Аннотация. В работе предложены модель текста и метод сравнения текстов, предназначенные для решения поисково-аналитических задач. В отличие от известных подходов метод сравнения текстов базируется на комплексном сопоставлении лексико-морфологической, синтаксической и семантической информации, автоматически устанавливаемой системами машинного анализа текстов. В работе рассмотрены вопросы применения предложенного метода для реализации полнотекстового, фразового, семантического и вопросно-ответного поиска как части сервисов поисково-аналитической системы.

Ключевые слова: полнотекстовый поиск, поисково-аналитические системы, оценка релевантности результатов поиска, ранжирование результатов поиска, фразовый поиск, вопросно-ответный поиск, семантический поиск, анализ научно-технической информации.

Введение

С ростом количества накопленной текстовой информации возникает потребность её систематизации для эффективного поиска и анализа. Задача информационного поиска, возникшая ещё в докомпьютерную эпоху в области библиотечного дела, стала одной из ключевых в информатике с развитием компьютерных технологий и Интернета [1, 2]. Неотъемлемой частью информационных систем, работающих с естественным языком (ЕЯ), являются технологии полнотекстового поиска, поскольку ЕЯ является основным средством коммуникации между людьми, и значительная часть накопленной информации представлена в виде ЕЯ-текстов.

Согласно ГОСТ 7.73-96, полнотекстовый поиск информации – «автоматизированный документальный поиск, при котором в качестве поискового образа документа используется его полный текст или существенные части текста». Поисковый запрос формулируется пользователем системы в виде совокупности ключевых слов [3], осмысленной фразы на ЕЯ или предложения (возможно, вопросительного) [4]. Поисковая машина выполняет поиск в собственной информационной базе и выдаёт результат в виде списка ссылок на документы, ранжированные в соответствии с некоторыми критериями [5]. Эти критерии предназначены для оценки степени соответствия найденных документов запросу пользователя – оценке релевантности результатов [6, 7].

Функции полнотекстового поиска реализованы в большинстве современных информационных систем: в поисковых машинах Интернета, таких как Яндекс, Гугл, Бинг и др., специализированных поисковых машинах (например, для поиска в текстовых базах данных), информационно-аналитических системах Scirus [8], Google Академия [9] и других.

Качество результатов поиска зависит от алгоритма ранжирования документов – способа автоматической оценки соответствия найденных текстов запросу пользователя. В поисковых машинах Интернета используется широкий круг критериев ранжирования документов, таких как ссылочное ранжирование – PageRank

[1] Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту № 07.514.11.4134 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

[13, 14], классификацию запросов по типам (транзакционный, информационный, навигационный и др.) [1], анализ поискового поведения пользователей [10, 15], учёт географического положения, анализ предпочтений пользователей на основе истории поиска и многие другие факторы [16–18]. Однако не все эти критерии применимы для решения задач поиска в аналитических системах и электронных библиотеках. В этих системах ключевую роль играет поиск документов с учётом метаинформации и метод оценки текстовой релевантности, т.е. соответствия найденного документа запросу пользователя, выраженному на ЕЯ в форме ключевых слов, фраз или предложений.

Решение задач полнотекстового поиска в области информационно-аналитической обработки текстов базируется на сопоставлении текстов и оценке их сходства путём сравнения различных признаков. Изначально основным методом сравнения текстов было их сопоставление по ключевым словам. Этот метод и соответствующие математические модели были созданы в середине XX века с развитием первых ЭВМ. Классические алгоритмы информационного текстового поиска рассматривают запрос пользователя как множество ключевых слов и упорядочивают документы, содержащие слова запроса, по убыванию информационной значимости этих слов в найденных текстах. Одной из моделей, предназначенных для оценки значимости слов ЕЯ в текстах, является модель Ципфа-Мандельброта [19, 20]. Эта модель получила широкое применение в области статистической лингвистики и информационного поиска в последней четверти XX века благодаря своей простоте и универсальности. Модель Ципфа-Мандельброта и её реализация в виде метода сравнения запроса и документов TF-IDF [21–23] оказалась достаточно эффективной для простого поиска по отдельным словам. Дальнейшее развитие этих статистических законов привело к появлению эмпирических формул семейства «Best Match» (BM-18, BM-25, BM25f) [24–26] и широкому распространению представления текстов в виде векторов ключевых слов.

Однако для поисковых запросов, сформулированных в виде предложений или фраз на ЕЯ, а также при вопросно-ответном и терминологическом поиске векторная модель далеко не всегда даёт качественные результаты. Причина этого состоит в том, что методы на основе векторной модели не учитывают информацию о вязях и отношениях слов в тексте. При этом качественное решение задач поисково-аналитической обработки, таких как автоматическое аннотирование текстов, выявление текстовых заимствований, вопросно-ответный поиск, невозможно без учёта лингвистической информации.

Современные методы информационного поиска и сравнения текстов учитывают следующие критерии при сопоставлении текстов друг с другом [1]:

1) совместную встречаемость, порядок слов запроса и расстояние между словами в текстах документов [27, 33];

2) синтаксические и семантические связи между словами запроса [29, 30];

3) предметную область и тематику текстов [11, 36, 39, 40];

4) отношения между словами текста в тезаурусе или онтологии предметной области (синонимические, родовидовые и др.) [37, 38].

Критерии первой группы относятся к нелингвистическим. Они не требуют проведения лингвистического анализа (за исключением морфологического, который применяется в настоящее время практически во всех системах информационного поиска для нормализации слов и разрешения омонимии).

Критерии второй группы основываются на результатах синтаксического и семантического анализа для установления соответствующих характеристик текстов [31]. Учёт лингвистической информации позволяет реализовать вопросно-ответный режим поиска [29, 30, 32] и способствует повышению точности сравнения текстов [34, 35].

Критерии третьей группы предполагают наличие таксономии (множества тематических рубрик), заданной априорно или построенной автоматически, например, путём кластеризации [11]. Сравнение текстов производится с учётом информации об их тематической принадлежности.

Критерии четвёртой группы требуют применения тезаурусов или лингвистических онтологий для установления связей между словами и терминами (концептами) предметных областей. При сравнении текстов учитываются связанные термины, встречающиеся в сопоставляемых текстах.

Таким образом, основной тенденцией в области методов поиска и анализа текстовой ин-

формации является учёт множества различных факторов с целью предоставления пользователю поисково-аналитических сервисов, обладающих высокой точностью и полнотой.

Целью настоящего исследования является разработка модели текста и метода комплексного сравнения текстов, основанных на признаках, получаемых при морфологическом, синтаксическом и семантическом анализе.

Предлагаемая модель является развитием предложенной в работах [48–50] реляционно-ситуационной модели (РСМ) Г.С. Осипова - Г.А. Золотовой, базирующейся на неоднородных семантических сетях [52]. Представленные в статье модель текстов и метод сравнения текстов являются результатами многолетней исследовательской работы, проводимой в ИСА РАН под руководством Г.С. Осипова в области интеллектуального семантического поиска [48, 49]. В качестве развития РСМ автором предложена её модификация, которая:

– учитывает теговую разметку текста для поиска с учетом метаинформации документов;

– вводит оценки информационной значимости вхождений слов в текстах;

– учитывает синтаксические структурные связи между вхождениями слов;

– учитывает возможную омонимию отдельных слов;

– вводит дополнительные признаки для формализации метода сравнения произвольных текстов по ряду критериев.

Автором сформулированы критерии сравнения текстов с учётом лексико-морфологической, синтаксической и семантической информации, и формализован общий метод сравнения текстов на основе этих критериев.

В заключительной части настоящей работы рассматриваются вопросы применения разработанного метода для решения поисково-аналитических задач, в частности, для оценки релевантности результатов информационного поиска.

1. Модель текста

Формализацию метода сравнения текстов начнём с описания модели текста, на которой базируется разрабатываемый метод. Для заданного текста на ЕЯ рассмотрим конечное множество вхождений слов $W=\{w_i\}$, из которых он состоит. Для описания информации, связанной с вхождениями слов, определим:

– Отображение из множества вхождений W во множество нормальных (словарных) форм лексем D: $\delta(w): W \to D$ – для определения нормальных форм вхождений слов.

– Отображение из множества вхождений W во множество форм лексем, Ф: $\varphi(w): W \to \Phi$ – для определения форм вхождений слов в текст.

– Конечное множество тегов (меток) $T=\{t_i\}$, и отображение $\tau(w): W \to T$, сопоставляющее каждому вхождению слова некоторый тег (соответствующий тегу гипертекстовой разметки или тегу разметки по метаданным, например, вхождению слова в заголовок текста, в список авторов).

– Числовую функцию, задающую вес вхождения слова в тексте: $v(w)$.

– Числовую функцию, задающую вес (значимость) тегов: $\omega(t)$.

Для учёта синтаксических зависимостей введём в рассмотрение следующие объекты:

– Множество типов синтаксических связей: $SR=\{np, vp\}$. Где np – обозначение для именной группы, vp – глагольной группы [42, 43]. В общем случае могут рассматриваться и другие синтаксические связи.

– Бинарное отношение на множестве вхождений слов: $\Sigma \subseteq W^2$, устанавливающее синтаксические связи между вхождениями слов согласно работам [42 – 45]. Будем считать, что в паре элементов (w_i,w_j) первый элемент w_i соответствует вхождению главного слова синтаксической группы, а второй элемент w_j – вхождению зависимого слова.

– Семейство отношений $Syntax = \left\{ \Sigma_z \subseteq \Sigma \mid z \in SR, \bigcup_{z \in SR} \Sigma_z = \Sigma \right\}$ – разбиение множества Σ, где каждое из отношений Σ_z определяет тип синтаксической связи $z \in SR$ для всех пар вхождений слов $(\mathrm{w_i}, \mathrm{w_j}) \in \Sigma_z$.

Для представления семантической информации в текстах рассмотрим:

– Конечное множество «семантических ролей» – значений синтаксем $Roles=\{abl, ag, ...\}$ в соответствии с работами [46, 47].

– Отображение $\rho(w_i):\widehat{W}\to 2^{Roles}$, где $\widehat{W}\subset W$ – подмножество вхождений слов текста, которые являются главными словами именных групп, составляющих множество синтаксем [48, 49]. Отображение ρ сопоставляет словам текста семантические значения соответствующих синтаксем [49, 50]. Образом элемента w_i является подмножество семантических значений из множества *Roles*. А полный образ отображения – некоторое подмножество булеана множества *Roles*. Таким образом, вхождение слова может иметь 0 и более семантических значений в тексте.

– Множество типов семантических связей *R*, устанавливаемых в соответствии с работами [50, 51] на множестве значений синтаксем.

– Бинарное отношение $\Omega\subseteq\widehat{W}^2$, определяющее семантически связанные вхождения слов в тексте.

– Семейство отношений $\mathrm{SemRels}=\left\{\Omega_x\subseteq\Omega \mid x\in R,\bigcup_{x\in R}\Omega_x=\Omega\right\}$ – разбиение множества Ω, где каждое отношение Ω_x определяет тип семантической связи $x\in R$ для всех пар вхождений слов $(\mathrm{w_i},\mathrm{w_j})\in\Omega_x$.

Семантическая структура текста в рассматриваемой модели полностью задаётся в виде двух множеств:

– *SemRoles*={<w_i,$\rho(w_i)$>} – главные слова именных групп в составе синтаксем и соответственные семантические значения;

– *SemRels* – разбиение множества Ω на классы, состоящие из пар вхождений слов, связанных семантически.

В представленной модели текст характеризуется множеством различных факторов:

– внетекстовых (теговая разметка, веса вхождений слов);

– лексико-морфологических (нормальные формы лексем, формы вхождений слов в текст);

– структурно-синтаксических (словосочетания, представленные именными и глагольными группами);

– семантических (значения синтаксем и семантические связи).

Для сопоставления текстов с учётом вышеописанных факторов введём в модель разбиение текста на предложения (возможно, сложные), заданное в виде множества $S=\{s_k\}$, где $\mathrm{s_k}=\{\mathrm{w_{i_k}},...,\mathrm{w_{j_k}}\}$, где $i_k,...,j_k$ – последовательность индексов вхождений слов в тексте, входящих в *k*-е предложение, а j_k-i_k – длина *k*-го предложения. Т.е. $\mathrm{S}\subset 2^{\mathrm{W}}$ (где 2^{W} – булеан множества *W*), причём *S* является разбиением (поскольку $\mathrm{s_k}\cap\mathrm{s_l}=\varnothing$ и $\forall\mathrm{w_j}\exists\mathrm{s_k}:\mathrm{w_j}\in\mathrm{s_k}$).

2. Метод сравнения текстов

Пусть задан эталонный текст τ^1 и сопоставляемый с ним текст τ^2. Введём функционал, оценивающий близость текстов τ^1 и τ^2 в рамках описанной модели.

Оценку близости текстов τ^1 и τ^2 будем проводить по множествам предложений этих текстов: S^1 и S^2 соответственно. Выберем два произвольных предложения $\mathrm{s_p^1}\in S^1$ и $\mathrm{s_r^2}\in S^2$. Рассмотрим множество $N\left(\mathrm{s_p^1},\mathrm{s_r^2}\right)=\left\{(w^1,w^2)\in W^1\times W^2 \mid w^1\in\mathrm{s_p^1}\ \&\ \exists w^2\in\mathrm{s_r^2}:\delta(w^1)=\delta(w^2)\right\}$ – суть множество пар соответственных вхождений слов предложения эталонного текста $\mathrm{s_p^1}$, совпадающих по нормальной форме лексемы с вхождениями слов в предложении $\mathrm{s_r^2}$ сопоставляемого текста. Сопоставление предложений производится путём сравнения соответственных вхождений слов в тексты, т.е. на элементах множества $N\left(\mathrm{s_p^1},\mathrm{s_r^2}\right)$.

Рассмотрим критерии сравнения предложений $\mathrm{s_p^1}$ и $\mathrm{s_r^2}$.

1. Покрытие предложения-эталона $\mathrm{s_p^1}$ предложением $\mathrm{s_r^2}$ определим как:

$$I_1\left(\mathrm{s_p^1},\mathrm{s_r^2}\right)=\sum_{(w^1,w^2)\in N\left(\mathrm{s_p^1},\mathrm{s_r^2}\right)}v(w^1). \qquad (1)$$

Здесь предполагается, что слова эталонного текста взвешены. В качестве функции для определения весов $v(w^1)$ может выступать классическая схема оценки информационной значимости на основе IDF [1, 2, 53] или характеристика тематической значимости (ХТЗ), описанная в работах [12, 54, 55]. Применение ХТЗ оправдано, если для эталонного текста τ^1 известна его тематическая принадлеж-

ность. Потребуем, чтобы сумма весов текста-эталона была равна единице:

$$\sum_{w^1 \in W^1} v(w^1) = 1. \qquad (2)$$

2. Общая оценка информационной значимости слов предложения-эталона s_p^1 в предложении s_r^2 текста определяется формулой:

$$I_2\left(s_p^1, s_r^2\right) = \sum_{(w^1,w^2)\in N\left(s_p^1,s_r^2\right)} f(w^1,w^2)v(w^2)v(w^1). \qquad (3)$$

Функционал $f(w^1,w^2)$ определяет «штраф» за несовпадение форм вхождений слов w^1,w^2:

$$f(w^1,w^2) = \begin{cases} 1, \varphi(w^1) = \varphi(w^2) \\ f_0, \textit{в противном случае} \end{cases},$$

где $f_0 \le 1$ – параметр метода.

В качестве весов $v(w^1)$ аналогично предыдущему пункту могут использоваться IDF или ХТЗ (если известна тематическая принадлежность текстов) с сохранением условия нормировки (2). В качестве функции для определения весов $v(w^2)$ может быть выбрана классическая схема TF [1, 2] или её модификации [23, 24, 26]. Дополнительным ограничением на величину $v(w^2)$ является следующее условие:

$$0 \le v(w^2) \le 1. \qquad (4)$$

В частности, в качестве $v(w^2)$ может рассматриваться модификация предложенной в работе [54] логарифмической оценки lTF:

$$v(w^2) = \omega\left(\tau(w^2)\right)\frac{\log\left(1+\sum\limits_{w'\in\left\{w''\in W \mid \delta(w^2)=\delta(w'')\right\}} \omega\left(\tau(w')\right)\right)}{\log\left(1+\sum\limits_{w'\in W}\omega\left(\tau(w')\right)\right)}. \qquad (5)$$

Логарифмическая оценка веса вхождения w пропорциональна произведению тегового веса, связанного с этим вхождением – $\omega(\tau(w))$ на величину, определяющую информационную значимость вхождения слова в тексте. Последняя величина, согласно работе [54], представляет собой разность между количеством информации, приходящимся на наиболее редко встречающуюся в тексте лексему, и частотой встречаемости лексемы $\varphi(w)$. В отличие от работы [54] предлагается оценивать эту величину не прямым подсчётом количества повторений лексемы в тексте, а суммировать соответственные теговые веса $\omega(\tau(w))$ – формула (5).

Заметим, что из условий (2) и (4) следует, что выполняется соотношение:

$$0 \le I_2\left(s_p^1, s_r^2\right) \le 1. \qquad (6)$$

3. Оценку сходства предложения-эталона s_p^1 и предложения s_r^2 текста на основе совпадения синтаксических структур зададим формулой:

$$I_3\left(s_p^1, s_r^2\right) = \frac{\sum\limits_{(w^1,w^2)\in N_{Syn}\left(s_p^1,s_r^2\right)} v(w^1)}{\sum\limits_{w^1\in\left\{w\in W^1 \mid \exists w'\in W^1:(w,w')\in\Sigma\right\}} v(w^1)}. \qquad (7)$$

В этой формуле множество $N_{Syn}\left(s_p^1, s_r^2\right) = \left\{(w^1,w^2)\in N\left(s_p^1,s_r^2\right) \mid \exists w_i^1 \in W^1, \exists w_j^2 \in W^2, \exists z \in SR : (w_i^1, w_j^2) \in N\left(s_p^1, s_r^2\right) \& (w^1, w_i^1) \in \Sigma_z^1 \;\&\; (w^2, w_j^2) \in \Sigma_z^2\right\}$ – суть множество пар соответственных вхождений слов эталонного предложения s_p^1 и вхождений слов сопоставляемого предложения s_r^2, для которых совпадают (по нормальным формам лексем) главные $(w^1,w^2)\in N\left(s_p^1,s_r^2\right)$ и зависимые $(w_i^1,w_j^2)\in N\left(s_p^1,s_r^2\right)$ слова, и эти слова образуют синтаксические группы одинакового типа. Знаменатель формулы (7) – есть совокупный вес вхождений слов, которые являются главными словами в синтаксических группах в эталонном тексте. Величина $I_3\left(s_p^1, s_r^2\right)$ – суть отношение суммарного веса вхождений слов, у которых совпадают синтаксические связи, к общему весу вхождений слов, имеющих в эталоне синтаксически связанные слова.

4. Сходство предложения-эталона s_p^1 и предложения s_r^2 текста на основе совпадения семантических значений определяется формулой:

$$I_4\left(s_p^1, s_r^2\right) = \frac{\sum\limits_{(w^1,w^2)\in N\left(s_p^1,s_r^2\right)} \left|\rho(w^1)\cap\rho(w^2)\right|}{\sum\limits_{w^1\in W^1}\left|\rho(w^1)\right|}. \qquad (8)$$

Числитель в формуле (8) выражает количество совпавших семантических значений у со-

ответственных слов предложения эталона и предложения сопоставляемого текста. Знаменатель формулы (8) задаёт условие нормировки на 1 по всем вхождениям слов, имеющим семантические значения (по эталону в целом).

5. Для оценки сходства предложения-эталона s_p^1 и предложения s_r^2 текста на основе совпадения семантических связей рассмотрим множество
$SemR(w) = \{h \in Roles \mid \exists w' \in W, \exists x \in R : (w, w') \in \Omega_x \,\&\, h \in \rho(w')\}$. Это – суть множество значений синтаксем, приписанных соответствующим вхождениям слов, которые связаны семантической связью с вхождением слова *w*. Тогда сходство предложения-эталона s_p^1 и предложения s_r^2 текста на основе семантических связей определим формулой:

$$I_5(s_p^1, s_r^2) = \frac{\sum_{(w^1,w^2)\in N(s_p^1,s_r^2)} \left|SemR(w^1) \cap SemR(w^2)\right|}{\sum_{w^1 \in W^1} \left|SemR(w^1)\right|}. \quad (9)$$

Аналогично формуле (8) числитель в формуле (9) выражает количество совпавших семантических связей у соответственных слов предложения эталона и предложения сопоставляемого текста. Знаменатель формулы задаёт условие нормировки на 1 по всем вхождениям слов, имеющих семантические связи в эталоне в целом.

Общая оценка близости предложения-эталона s_p^1 и предложения s_r^2 сопоставляемого текста определяется взвешенной суммой вышеописанных величин:

$$Sim(s_p^1, s_r^2) = \sum_{n=1}^{5} \lambda_n I_n(s_p^1, s_r^2). \quad (10)$$

В формуле (10) набор $L=\{\lambda_i | i=1..5\}$ задаёт набор параметров метода, определяющих вклад каждого из критериев оценки близости в итоговую величину. Зададим ограничение на набор параметров *L*:

$$\sum_{n=1}^{5} \lambda_n = 1. \quad (11)$$

Теперь из всех возможных предложений в сопоставляемом тексте τ^2 выберем те, которые наилучшим образом (в смысле максимизации оценки (10)) соответствуют предложению эталона s_p^1. Величина, определяющая наибольшие оценки близости предложений найденного текста к предложению эталона s_p^1, задаётся следующей формулой:

$$Y(s_p^1, \tau^2) = \max_{s_r^2 \in S^2} \{Sim(s_p^1, s_r^2)\}. \quad (12)$$

Общая оценка близости текста-эталона τ^1 по отношению к тексту τ^2 выражается следующей формулой:

$$X(\tau^1, \tau^2) = \sum_{s_p^1 \in S^1} Y(s_p^1, \tau^2). \quad (13)$$

Поскольку эталонный и сопоставляемый тексты разделяются на предложения путём построения разбиения, а также в силу предложенных условий нормировки (2), (4), (6), (11), нормированности величин (7) – (9) и построению формул (10) – (13) нетрудно показать, что верно следующее утверждение: $0 \le X(\tau^1, \tau^2) \le 1$. Это свойство позволяет рассматривать величину, заданную формулой (13), как оценку релевантности запроса пользователя (эталонный текст) и найденных текстов. Если оценка близости не превышает минимального порога X_{min} (параметр алгоритма) для некоторого текста, то этот текст можно считать нерелевантным.

3. Применение метода для решения поисково-аналитических задач

Представленный в работе метод сравнения текстов реализован в системе поиска и анализа научно-технической информации. Реализация полнотекстового поиска включает следующие алгоритмы:

- поиск по ключевым словам и словосочетаниям с учётом синтаксиса;
- фразовый поиск;
- семантический вопросно-ответный поиск;
- аннотирование результатов поиска.

Рассмотрим далее некоторые практические аспекты реализации функций полнотекстового поиска на основе разработанного метода сравнения текстов.

1. При поиске запрос пользователя на ЕЯ рассматривается в качестве эталонного текста.

По нормальным формам вхождений слов запроса производится выборка данных из специальной структуры данных – инвертированного поискового индекса (ИПИ), содержащего информацию о вхождениях слов в тексты документов. Отметим, что запрос может состоять как из нескольких предложений или фраз, так и из одного слова. В случае коротких запросов морфологический анализатор не способен разрешить омонимию (на уровне нормальных форм) и выдаёт несколько вариантов нормализации. Чтобы обеспечить полноту результатов поиска и не потерять релевантные результаты, выборка информации из ИПИ производится с учётом омонимов. Предложенная модель текста незначительно модифицируется, чтобы учесть этот факт: вместо отображения $\delta(w): W \to D$ рассматривается отображение $\delta'(w): W \to 2^D$ (т.е. вхождение слова может иметь 1 или более вариант нормализации). При этом в определении множества $N\left(\mathrm{s}_\mathrm{p}^1, \mathrm{s}_\mathrm{r}^2\right)$ вместо условия $\delta(w^1) = \delta(w^2)$ рассматриваем условие $\delta(w^1) \cap \delta(w^2) \neq \varnothing$. Все остальные формулы остаются прежними.

2. При ранжировании результатов поиска документы упорядочиваются по убыванию величины $X(\tau^1, \tau^2)$. По умолчанию считается, что запросу релевантны все документы, содержащие хотя бы одно слово запроса. В булевом поиске [1, 2] этому принципу соответствует объединение слов запроса через оператор «ИЛИ». Однако нерелевантные результаты можно исключить из выдачи по заданному порогу минимального сходства X_{min}. Реализация полнотекстового поиска поддерживает специализированный язык «пользовательской разметки» запросов, позволяющий:

2.1. Исключать из результатов поиска документы, содержащие выбранные слова запроса (оператор «\»);

2.2. Исключать из рассмотрения отдельные предложения документов, содержащие выбранные слова запроса (оператор «~»);

2.3. Задавать слова запроса, которые обязательно должны присутствовать в результатах поиска (оператор «+»), в том числе в определённой форме (оператор «&»).

3. Поиск с учётом метаданных позволяет находить тексты, в которых определённые фрагменты (предложения) поискового запроса выделены специальными тегами гипертекстовой разметки либо тегами разметки метаданных, например, представляют собой заголовок текста, фамилии авторов и т.п. При поиске с учётом метаданных в определение множества $N\left(\mathrm{s}_\mathrm{p}^1, \mathrm{s}_\mathrm{r}^2\right)$ дополнительно через конъюнкцию добавляется условие $\tau(w^1) = \tau(w^2)$ – совпадение тегов у соответственных вхождений слов. Все остальные формулы остаются без изменений.

4. Фразовый поиск позволяет находить предложения, содержащие заданные слова запроса, связанные теми же синтаксическими связями, что присутствуют в запросе. Под фразой понимается некоторая часть предложения как совокупность вхождений слов, связанных друг с другом синтаксически. Таким образом, в качестве фразы может выступать термин предметной области («*интеллектуальные методы поиска*»), ФИО персоны («*Пётр Струве*») или организации, а также предложение целиком («*лиса ест манную кашу*») – любые конструкции, в которых синтаксический анализатор устанавливает связи между входящими в них словами. В отличие от функции поиска по точному соответствию, реализованной в большинстве поисковых машин, при фразовом поиске будут найдены предложения, соответствующие запросу лексически и синтаксически, но, возможно, отличающиеся порядком и формой вхождения слов («*интеллектуальный метод информационного поиска*», «*Пётр Бернгардович Струве*», «*кашу манную лисы едят с удовольствием*») – соответствующие синтаксические деревья приведены на Рис.1. При этом нерелевантные результаты («*поиск интеллектуальных методов анализа*») будут ранжированы ниже релевантных, поскольку им соответствуют синтаксические деревья другой структуры. Эти результаты могут быть полностью исключены из выдачи.

Для реализации фразового поиска модель текста дополняется разбиением на множество фраз $\Pi = \{\pi_{p,m}\}$, где $\pi_{p,\mathrm{m}} = \{\mathrm{w}_{\mathrm{i\,m}}, \ldots, \mathrm{w}_{\mathrm{j}_m}\}$, $i_m, \ldots j_m$ – последовательность индексов вхождений слов в предложении s_p, входящих в m-ю фразу: $\pi_{\mathrm{p,m}} \cap \pi_{\mathrm{q,n}} = \varnothing$ и $\forall \mathrm{w}_\mathrm{j} \exists! \pi_{\mathrm{p,k}} : \mathrm{w}_\mathrm{j} \in \pi_{\mathrm{p,k}}$ (каждая фраза содержит не менее одного слова, фразы не пересекаются). Разбиение на фразы строится внутри предложений. В тривиальном случае

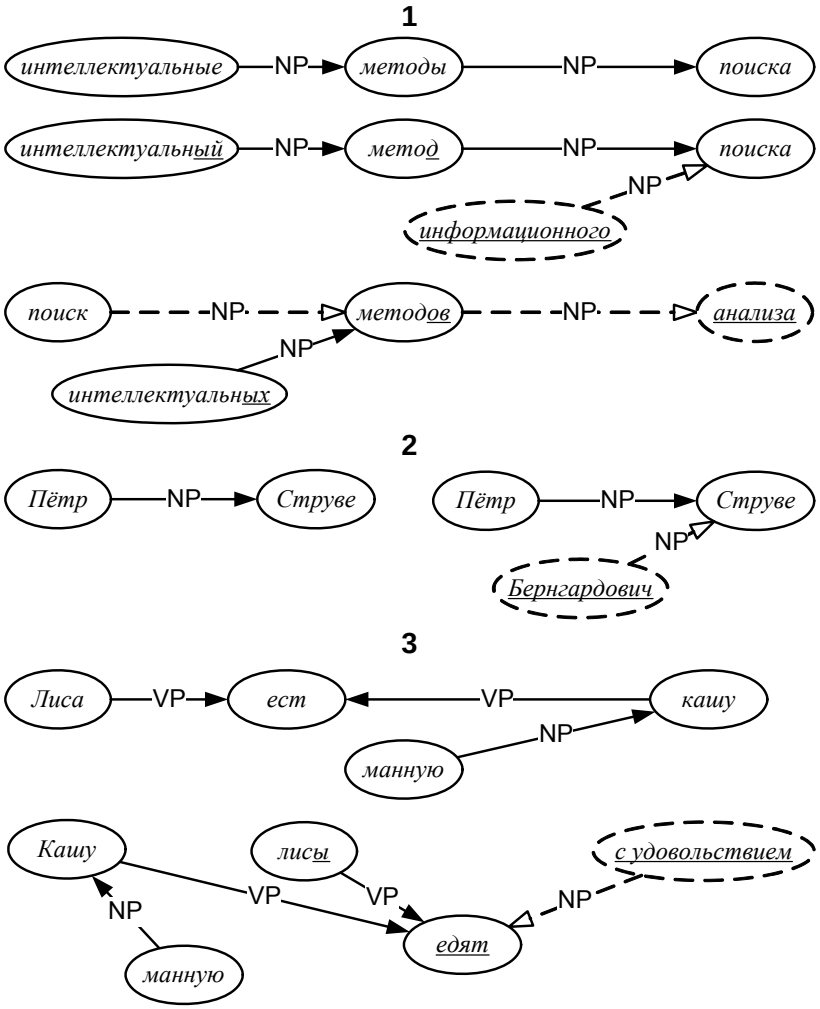

Рис. 1. Примеры синтаксических деревьев, сравниваемых при фразовом поиске

каждая фраза содержит в точности по одному слову предложения. К фразам применимы вышеперечисленные операции языка запросов 2.1–2.3; для объединения слов во фразы при этом применяются фигурные скобки «{…}». При фразовом поиске на этапе выборки данных из ИСИ предложение найденного документа сопоставляется с предложением запроса таким образом, чтобы в них совпала хотя бы одна из фраз. Под совпадением фраз здесь подразумевается совпадение всех вхождений слов по нормальным формам, а также совпадение всех синтаксических связей у соответствующих вхождений слов. Т.е. $N'_{Syn}\left(\pi^1_{\text{п}},\pi^2_{\text{г}}\right)=\left\{(w^1,w^2)\in N \mid \exists w^1_i:(w^1,w^1_i)\in\Sigma^1\ \&\ \exists w^2_j:(w^2,w^2_j)\in\Sigma^2\ \&\ (w^1_i,w^2_j)\in N\left(\pi^1_{\text{п}},\pi^2_{\text{г}}\right)\&\ \sigma(w^1,w^1_i)=\sigma(w^2,w^2_j)\right\}$, и при этом требуется, чтобы для $\forall w^1\in\pi^1_p:\exists(w^1,\hat{w}^1)\in\Sigma^1$ было выполнено: $\exists(w^1,w^2)\in N'_{Syn}\left(\pi^1_{\text{п}},\pi^2_{\text{г}}\right)$.

5. Для учёта тезаурусных связей и реализации поиска с учётом синонимичных или родовидовых отношений применяется перифраз исходного текста. Во множество слов W^1 текста запроса (и в соответствующие предложения) добавляются слова-синонимы. В общем случае синоним подбирается не к отдельному слову, а к словосочетанию, составляющему фразу (соответствующую некоторому термину предметной области, например, «*поисково-аналитическая система*» – «*система поиска и анализа научной информации*»). Для реализации процедуры пополнения исходного текста в предложенную модель вводится бинарное

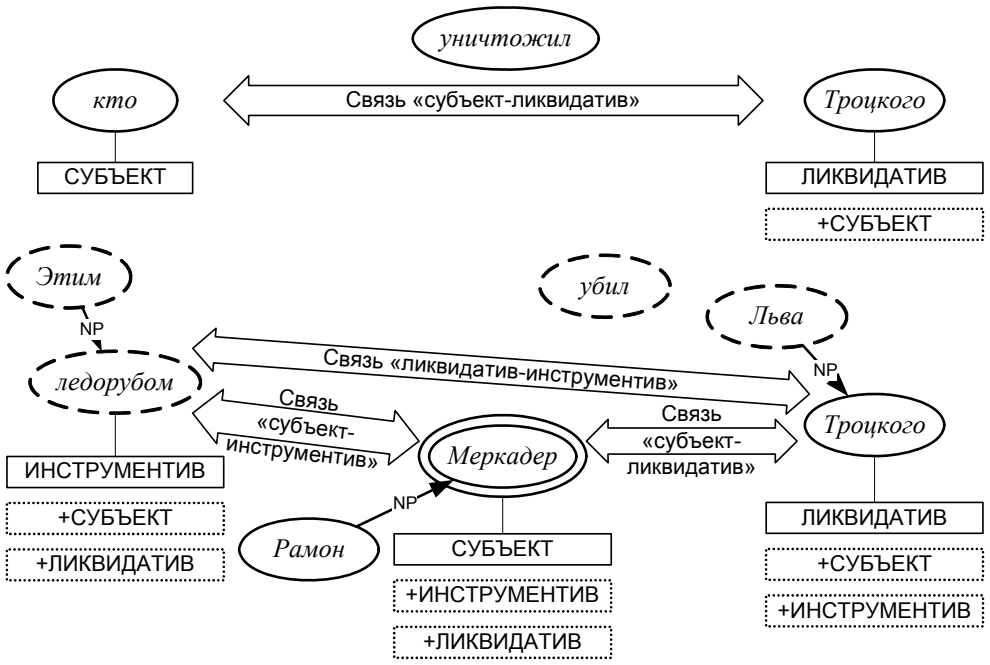

Рис. 2. Пример сравнения семантико-синтаксических структур вопроса и найденного ответа

отношение на множестве фраз $P \subset \Pi^2$, причём $(\pi^1_{\mathrm{p}}, \widetilde{\pi}^1_{\mathrm{q}}) \in P$ тогда и только тогда, когда фраза $\widetilde{\pi}^1_{\mathrm{q}}$ является концептом тезауруса (добавленным в текст), связанным с фразой запроса π^1_{q}. В остальном процедура поиска не отличается от вышеизложенного фразового поиска за исключением того, что слова добавленных фраз не изменяют условий нормировки исходного текста и, возможно, рассматриваются с уменьшенным весом: $\sum_{w^1 \in \widetilde{\pi}^1} \nu(w^1) \le \sum_{w^1 \in \pi^1} \nu(w^1)$.

6. Вопросно-ответный и семантический алгоритмы поиска является дальнейшим развитием алгоритма фразового поиска. Если поисковый запрос пользователя представляет собой осмысленное предложение или вопрос, то в нём устанавливается семантическая информация, которая позволяет находить предложения, близкие по смыслу исходному, или ответ на поставленный вопрос. Пример сравнения семантики вопроса пользователя («*кто уничтожил Троцкого?*») и найденного по слову «*Троцкий*» ответа («*Этим ледорубом Рамон Меркадер убил Льва Троцкого*») приведён на Рис.2. Вопросительное слово исключается из рассмотрения, и поиск проводится только по оставшимся словам с учётом семантической информации.

Предложенный в настоящей работе метод позволяет проводить сравнение текстов с учётом различных лингвистических признаков. В зависимости от решаемой задачи (классический или фразовый поиск, семантический поиск близких ситуаций или вопросно-ответный поиск) значимость критериев и их вклад в итоговую оценку сходства текстов может варьироваться. Вклад каждого из критериев определяется конкретными значениями в наборе параметров (L, f_0, X_{min} и др.). Настройка метода под каждую конкретную задачу должна проводиться на основе контрольных выборок с подбором таких параметров, которые обеспечивают наилучшее качество решения.

Заключение

Представленный в работе аппарат реализован в виде алгоритмического и программного обеспечения поисково-аналитической машины Exactus Expert [48], разработанной в ИСА РАН. Предложенный метод сравнения текстов, базирующийся на исследованной модели текстов, использует лексико-морфологические, синтаксические и семантические критерии для оценки сходства текстов. На применении разработанного метода основывается реализация сервисов полнотекстового семантического, фразового и

вопросно-ответного поиска. Полнотекстовый семантический поиск обеспечивает доступ пользователей к интересующей их научно-технической информации. Режим фразового поиска позволяет экспертам-пользователям поисково-аналитической системы Exactus Expert с высокой точностью проводить тематический анализ публикаций и отслеживать динамику употребления терминов предметной области в коллекциях научно-технических документов за различные периоды времени. Помимо ранжирования документов в поисковой выдаче результаты работы метода применяются для аннотирования результатов поиска (построения поисковых сниппетов).

Алгоритмы решения поисково-аналитических задач, базирующиеся на разработанном методе сравнения текстов и модели текстов, были многократно апробированы на семинарах РОМИП [28, 32, 41] в 2006–2010 годах. По независимым экспертным оценкам РОМИП разработанные поисково-аналитические алгоритмы имеют высокие показатели качества (по полноте, точности и другим метрикам).

Предметом дальнейших авторских исследований является разработка метода поиска недобросовестных текстовых заимствований в научно-технических и учебных документах, с использованием изложенного метода сравнения текстов и его реализация в виде сервиса поисково-аналитической машины.

Литература

1. Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.
2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011. — ISBN 978-5-8459-1623-5.
3. Spink, A.; Wolfram, D.; Jansen, M. B. and Saracevic, T. Searching the web: The public and their queries. Journal of the American society for information science and technology, Wiley Online Library, 2000, №52, P.P. 226-234.
4. Lehnert, W. G. Grosz, B. J.; Sparck Jones, K. and Webber, B. L. (Eds.) A Conceptual Theory of Question Answering. Natural Language Processing, Kaufmann, 1986, P.P. 651-657.
5. Golliher, S. A. Search engine ranking variables and algorithms. SEMJ. ORG, 2008, №1 Supplemental Issue
6. Mizzaro, S. Relevance: The whole history. Journal of the American society for information science, 1997, №48, P.P. 810-832
7. Golliher, S.A. Search Engine Ranking Variables and Algorithms. semj.org, 2008, №1, Supplemental Issue.
8. Scirus. / Сайт системы поиска по научным ресурсам Scirus. [Электронный ресурс] – URL: http://www.scirus.com/srsapp/, (дата обращения 14.03.2013).
9. Google Академия. / Сайт Google Академия. [Электронный ресурс] – URL: http://scholar.google.ru/ (дата обращения 14.03.2013).
10. Kromer, P.; Snasel, V.; Platos, J.; Owais, S.S.J. Implicit User Modelling for Web Search Improvement. // Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on. pp.309-314. doi: 10.1109/ISDA.2007.5.
11. Myonggwon Hwang; Hyunjang Kong; Sunkyoung Baek; Pankoo Kim. TSM. Topic Selection Method of Web Documents. // Modelling & Simulation, 2007. AMS '07. First Asia International Conference on , vol., no., pp.369,374, doi: 10.1109/AMS.2007.108.
12. Р.Е. Суворов, И.В. Соченков. Определение связанности научно-технических документов на основе характеристики тематической значимости. // Искусственный интеллект и принятие решений. М.: ИСА РАН, 2013, №1, с.33-40.
13. Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems, Elsevier, 1998, №30, P.P. 107-117.
14. Page, L.; Brin, S.; Motwani, R. and Winograd, T. The PageRank citation ranking: bringing order to the web. Stanford InfoLab, 1999.
15. С.Протасов Новое ранжирование Рамблера, Почему мы отказались от MatrixNet, Ноябрь 2010 [Электронный ресурс] URL: http://slashzone.ru/parser/Protasov-Rambler-NR2010.pdf (дата обращения 8.03.2013).
16. Компания Яндекс – Матрикснет // 2009 [Электронный ресурс] URL: http://company.yandex.ru/technologies/matrixnet/index.xml (дата обращения 29.10.2012)
17. И. Зябрев, О. Пожарков. Жадные алгоритмы в Яндексе. // 2010 [Электронный ресурс] URL: http://www.altertrader.com/publications20.html (дата обращения 23.03.2013).
18. Paananen, A. Comparative Analysis of Yandex and Google Search Engines. //.Metropolia Ammattikorkeakoulu. // Master's thesis. 2012 [Электронный ресурс] URL: http://publications.theseus.fi/handle/10024/46483 (дата обращения 23.03.2013).
19. Zipf, G. K. Selected studies of the principle of relative frequencies of language // Cambridge, Massachusetts: Harvard Unive, 1932.
20. Montemurro, M. A. Beyond the Zipf-Mandelbrot law in quantitative linguistics. // Physica A Statistical Mechanics and its Applications, 2001, №300, P.P. 567-578.
21. Jones, K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of documentation, MCB UP Ltd, 1972, №28, P.P. 11-21.
22. Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. // Journal of documentation, Emerald Group Publishing Limited, 2004, №60, P.P. 493-502.
23. Joachims, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. // DTIC Document, 1996.

24. Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. and Gatford, M. Okapi at TREC-3 // Proceedings of the Third Text REtrieval Conference TREC, 1994.
25. Zaragoza, H.; Craswell, N.; Taylor, M.; Saria, S. and Robertson, S. Microsoft Cambridge at TREC-13: Web and HARD tracks Proceedings of, 2004.
26. Joaquin Perez-Iglesias. Integrating the Probabilistic Model BM25: BM25F into Lucene, 2009 [Электронный ресурс] URL: http://nlp.uned.es/~jperezi/Lucene-BM25/ (дата обращения 23.03.2013).
27. Google Guide. Interpreting Your Query. [Электронный ресурс] URL: http://www.googleguide.com/interpreting_queries.htm (дата обращения 23.03.2013).
28. А.А. Киселев, И.В. Смирнов, И.А. Тихомиров, И.В. Соченков. Система интеллектуального поиска и анализа информации Exactus на РОМИП-2010 // Труды российского семинара по оценке методов информационного поиска РОМИП'2010. - Казань: Казан. ун-т, 2010. С49-69.
29. Shen, Dan, and Mirella Lapata. Using Semantic Roles to Improve Question Answering // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007 [Электронный ресурс] URL: http://www.aclweb.org/anthology/D/D07/D07-1002 (дата обращения 23.03.2013).
30. Stenchikova, S.; Hakkani-Tur, D.and Tur, G. QASR: Spoken Question Answering Using Semantic Role Labeling. // ASRU-2005, 9th biannual IEEE workshop on Automatic Speech Recognition and Understanding, 2005.
31. Daniel Gildea and Daniel Jurafsky. Automatic Labeling of Semantic Roles. // In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pp. 512–520, Hong Kong, October 2000.
32. Тихомиров И. А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus. // Труды четвертого российского семинара по оценке методов информационного поиска РОМИП'2006. Санкт-Петербург: НУ ЦСИ, 2006. - с. 80-85.
33. Chi Tian; Tezuka, T.; Oyama, S.; Tajima, K.; Tanaka, K. Web Search Improvement Based on Proximity and Density of Miltiple Keywords. // Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on , vol., no., pp.x133,x133, 2006. doi: 10.1109/ICDEW.2006.164.
34. Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. М.: ИСА РАН – №2, 2008. С. 3-10.
35. Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Olga Vybornova. Technologies for Semantic Analysis of Scientific Publications. //Proceedings of 2012 IEEE 6th International Conference Intelligent Systems (Yager, R.R., V. Sgurev and M. Hadjiski, Eds.), Vol. II, pp. 58-62.
36. Lin-Tao Lv; Li-Ping Chen; Hong-Fang Zhou. An improved topic relevance algorithm for vertical search engines. // Wavelet Analysis and Pattern Recognition, 2008. ICWAPR '08. International Conference on, vol.2, no., pp.753, 757, 30-31 Aug. 2008. doi: 10.1109/ICWAPR.2008.4635878.
37. Tao Cheng, Hady W. Lauw, Stelios Paparizos. Entity Synonyms for Structured Web Search. // IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 10, pp. 1862-1875, Oct. 2012, doi:10.1109/TKDE.2011.168.
38. Xing Wei, Fuchun Peng, Huihsin Tseng, Yumao Lu, Xuerui Wang, Benoît Dumoulin. Search with Synonyms: Problems and Solutions. // COLING (Posters). 2010, pp.1318-1326.
39. Lewandowski, Dirk. Query types and search topics of German Web search engine users. // Information Services & Use, 2006, vol. 26, n. 4, pp. 261-269.
40. Yue Wang , Hongsong Li , Haixun Wang , Kenny Q. Zhu. Toward Topic Search on the Web. // In the Proceedings of ER 2012.
41. Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И. А. Результаты и перспективы поискового алгоритма Exactus. // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. Санкт-Петербург: НУ ЦСИ, 2008. - С. 66-76.
42. А.Сокирко. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) // Дисс канд.т.н. [Электронный ресурс] - Режим доступа: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html, свободный. Проверено 23.01.2013.
43. Head-Driven Phrase Structure Grammar. // [Электронный ресурс] URL: http://hpsg.stanford.edu/ (дата обращения 23.03.2013).
44. Автоматическая Обработка Текста (АОТ). // [Электронный ресурс] - Режим доступа: http://www.aot.ru, свободный. Проверено 23.01.2013.
45. Link Grammar // [Электронный ресурс] URL: http://www.link.cs.cmu.edu/link/ (дата обращения 23.03.2013).
46. Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. – М. 2004. – 544 с.
47. Осипов Г.С. Методы искусственного интеллекта. – М.: ФИЗМАТЛИТ, 2011. – 296 с.
48. Osipov, G.; Smirnov, I.; Tikhomirov, I. and Shelmanov, A. Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications. // In Proceedings of the Workshop on Integrating IR technologies for Professional Search Moscow, Russian Federation, March 24, 2013, p.57-64. [Электронный ресурс] URL: http://ceur-ws.org/Vol-968/irps_10.pdf (дата обращения 23.03.2013).
49. Osipov, G.; Smirnov, I.; Tikhomirov, I.; Zavjalova O. Application of linguistic knowledge to search precision improvement. // InProceedings of 4th International IEEE conference on Intelligent Systems. Vol. 2, pp. 17-2–17-5, 2008.
50. G. Osipov. Methods for extracting semantic types of natural language statements from texts. // In 10th IEEE International Symposium on Intelligent Control, Monterey, California, USA, 1995.
51. Осипов Г.С. Приобретение знаний интеллектуальными системами. // М.: Наука. Физматлит, 1997.
52. Osipov, G. Formulation of Subject Domain Models: Part 1. Heterogeneous Semantic Nets. // Journal of Computer and Systems Sciences International. Scripta Technica Inc., New-York, 1992.

53. S.E. Robertson, C.J. van Rijsbergen and P.W. Williams. Probabilistic models of indexing and searching. In R.N. Oddy // Information Retrieval Research, pp. 35-56, London, 1981. Butterworths. [Электронный ресурс] URL: http://www.soi.city.ac.uk/~ser/papers/Robertson_vanRijsbergen_Porter.pdf (дата обращения 23.03.2013).
54. Э. Мбайкоджи, А. А. Драль, И. В. Соченков. Метод автоматической классификации коротких текстовых сообщений. // Информационные технологии и вычислительные системы. – 2012. – №3. – с.93 – 102.
55. Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке. // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2008. Т. 6. № 2. С.94-100.

Соченков Илья Владимирович. Инженер-исследователь ИСА РАН, научный сотрудник ООО "ТСА", ассистент кафедры информационных технологий РУДН. Окончил Российский университет дружбы народов в 2009 году. Автор 25 печатных работ. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, защита сетей, контентная фильтрация, компьютерная лингвистика. E-mail: sochenkov@isa.ru